SciDataCon 2016

Conference paper

Developing Criteria to Establish Trusted Digital Repositories.

John Faundeen, U.S. Geological Survey

Abstract

This presentation details the drivers, the process, and the outcomes of the U.S. Geological Survey's quest to establish criteria by which to judge it's own digital research resources as Trusted Digital Repositories. Drivers include recent U.S. legislation focused on agencies spending $100M USD or more annually on research activities. The process entailed searching for existing criteria sets from national and international organizations such as ISO, the U.S. Library of Congress, and the Data Seal of Approval. Complexity, cost, and usability were key discussion elements. The selected outcome was chosen that allows the process to be transparent, understandable, and defensible. Those factors are critical when judging between competing, internal units. Implementing the chosen criteria involved establishing a cross-agency team that interfaced with many levels throughout the organization.

Background

As the Nation's largest water, earth, and biological science and civilian mapping agency, the U.S. Geological Survey (USGS) collects, monitors, analyzes, and provides scientific understanding about natural resource conditions, issues, and problems. The diversity of our scientific expertise enables us to carry out large-scale, multi-disciplinary investigations and provide impartial scientific information to resource managers, planners, and other customers.

Drivers

On February 22, 2013, the U.S. Office of Science and Technology Policy (OSTP) issued a memorandum, *Increasing Access to the Results of Federally Funded Scientific Research*, which called on all Federal agencies with annual research and development (R&D) expenditures of more than $100 million to develop a plan to increase public access to the direct results of federally funded scientific research, including specifically peer-reviewed publications and digital data. The Department of the Interior's total annual Research and Development budget in FY 2015 was approximately $925 million. 74 percent ($686 million) of that funding was allocated to the USGS. The USGS Plan focused specifically upon the USGS's 'public access' activities, policies, and plans, as they affect both intramural and extramural research and development activities.

On May 9, 2013, the U.S. Office of Management and Budget (OMB) also released Memorandum M-13-13, *Open Data Policy--Managing Information as an Asset*. Individually and collectively these directives established the mandates for the U.S. Federal Government to transform data and information into useable and accessible digital artifacts and promote

and accelerate their release, subject to certain limitations imposed by privacy, confidentiality, and national security considerations.

Since the inception of USGS in 1879, the agency has maintained comprehensive internal and external policies and procedures for ensuring the quality and integrity of its science. This has led to the reputation of USGS being noted for science excellence and objectivity. In 1993 the first internal policies were instituted requiring preservation of digital assets. In 2003 the USGS establish the web-based USGS Publications Warehouse, its first digital library. In 2006, the scientific policies and procedures were updated, and are now known as USGS Fundamental Science Practices (FSP), a set of consistent practices, philosophical premises, and operational principles to serve as the foundation for research and monitoring activities related to USGS science. In January 2009, the Director of the USGS announced the establishment of a Fundamental Science Practices Advisory Committee (FSPAC). The function of the FSPAC is to address pending and new FSP issues, including previously unresolved issues, listen to questions and concerns about FSP from scientists and managers, and develop recommendations for resolving issues. The FSPAC serves as a sounding board for FSP issues and as a resource to USGS management and scientists by offering recommendations and guidance on planning and conducting scientific research and review, approval, and release processes to help ensure that the USGS continues to produce high quality, objective science information products. In 2012 the FSPAC established a Data Preservation Sub-Committee to provide recommendations to help identify and resolve USGS science data stewardship, preservation, and documentation issues which resulted in the 2015 USGS policy entitled “Fundamental Science Practices: Preservation Requirements for Digital Scientific Data”

Sources Reviewed

The Subcommittee also assists in the formulation of best practices and potential future FSP policy related to ensuring that USGS science data assets are preserved, available, and usable. In this capacity the sub-committee evaluated several existing criteria sets hoping to identify elements that would form our own review essentials related to *Trusted Digital Repositories*.

The first criteria set reviewed was the U.S. Federal RIM Program Maturity Model. Elements such as Strategic Planning, Leadership and Management, Resources, Policy, Standards, and Governance Framework plus Compliance Monitoring, Risk Management, Lifecycle Management, Retrieval and Accessibility, Security, and Protection were included in this work.

Another criteria set was compiled by the Digital Curation Centre entitled, “Where to keep research data: DCC Checklist for Evaluating Data Repositories.” This checklist was built around the following questions:

Is a reputable repository available?

Will it take the data you want to deposit?

Will it be safe in legal terms?

Will the repository sustain the data value?

Will it support analysis and track data usage?

The third approach originated from the U.S. National Oceanic and Atmospheric Administration. A paper entitled, “A Unified Framework for Measuring Stewardship Practices Applied to Digital Environmental Datasets” describes this method. The key components include Preservability, Accessibility, Usability, Production Sustainability, Data Quality Assurance, Data Quality Control/Monitoring, Data Quality Assessment, Transparency/Traceability and Data Integrity.

The Data Seal of Approval (DSA) approach involves addressing several questions including the following:

- The data producer provides the data together with the metadata requested by the data repository.
- The data repository has an explicit mission in the area of digital archiving and promulgates it.
- The data repository applies documented processes and procedures for managing data storage.
- The data repository has a plan for long-term preservation of its digital assets.
- Archiving takes place according to explicit work flows across the data life cycle.
- The data repository assumes responsibility from the data producers for access and availability of the digital objects.
- The data repository ensures the integrity of the digital objects and the metadata.
- The data repository ensures the authenticity of the digital objects and the metadata.
- The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

The International Standards Organization (ISO) has issued a standard labeled 16363-2012 related to records management. Section 3 focuses on Organizational Infrastructure related to trusted digital repositories. Items like governance, organizational viability, preservation policy framework, and financial sustainability are covered there. The 4th Section involves criteria on Digital Object Management. This segment details how ingest, preservation planning, and access management are handled. Section 5 deals with Infrastructure and Security Risk Management.

The U.S. Library of Congress sponsored National Digital Stewardship Alliance (NDSA) developed a “…tiered set of recommendations for how organizations should begin to build or enhance their digital preservation activities. A work in progress by the NDSA, it is intended to be a relatively easy-to-use set of guidelines useful not only for those just beginning to think about preserving their digital assets, but also for institutions planning the next steps in enhancing their existing digital preservation systems and workflows. It allows institutions to assess the level of preservation achieved for specific materials in their custody, or their entire preservation infrastructure. It is not designed to assess the robustness of digital preservation

programs as a whole since it does not cover such things as policies, staffing, or organizational support. The guidelines are organized into five functional areas that are at the heart of digital preservation systems: storage and geographic location, file fixity and data integrity, information security, metadata, and file formats."

The USGS Data Preservation Sub-Committee built upon the NDSA recommendations and replaced some text such as *fixity* to *checksums* to be more understandable to our agency personnel. The NDSA primary elements include the areas of Storage and Geographic Location, Data Integrity, Information Security, Metadata, and File Formats. The USGS also added the element of Physical Media because of the large role media decisions can have on the preservation of agency science data.

Path Chosen

After many discussions detailing the pros and cons of the various items and approaches used to assemble a criteria set USGS could use for determining our own *Trusted Digital Repositories*, the Data Preservation Sub-Committee recommended using a version of the Data Seal of Approval approach. In February of 2016 the Data Seal of Approval and the International Council of Scientific Unions (ICSU) World Data System (WDS) released a combined criteria set entitled, "DSA-WDS Partnership Working Group Catalogue of Common Requirements." The 16 primary elements in this new approach include addressing the following statements:

> The repository has an explicit mission to provide access to and preserve data in its domain.
>
> The repository maintains all applicable licenses covering data access and use and monitors compliance.
>
> The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.
>
> The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
>
> The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.
>
> The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).
>
> The repository guarantees the integrity and authenticity of the data.
>
> The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
>
> The repository applies documented processes and procedures in managing archival storage of the data.
>
> The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Archiving takes place according to defined workflows from ingest to dissemination.

The repository enables users to discover the data and refer to them in a persistent way through proper citation.

The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Implementation Strategy

The USGS established a cross-agency team to develop a strategic approach to meeting the required criteria. The Team will capitalize on the transparency of using an international criterion set developed by authoritative sources that provides a means by which agency facilities can be judged.

Summary

Several new data management policies have been developed and implemented recently. The establishment of criteria enabling the certification of agency *Trusted Digital Repositories* was the last remaining requirement. The adoption of DSA-WDS Partnership Working Group Catalogue of Common Requirements completes the lifecycle approach USGS has adopted to create, maintain, make accessible and preserve its scientific endeavours.

Acknowledgements

The author would like to thank Keith Kirk and Keith Richmond for their contributions.

Competing Interests

The author declares that he has no competing interests.

Notes

1 The USGS implementation of the criteria to judge Trusted Digital Repositories is expected to proceed for some time in the future. It may also evolve as lessons are learned.

References

Data Archiving and Networked Services 2016 Data Seal of Approval: On-line assessment tool. Netherlands. Available at http://www.datasealofapproval.org/en/information/guidelines/ [Last accessed 29 April 2016].

Data Seal of Approval-World Data System 2016 DSA-WDS Partnership Working Group Catalogue of Common Requirements. Research Data Alliance. Available at https://rd-alliance.org/groups/repository-audit-and-certification-dsa–wds-partnership-wg.html [Last accessed 29 April 2016].

International Standards Organization 2012 ISO 16363-2012: AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES. Available at http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=56510 [Last accessed 29 April 2016].

Joint Working Group of the Federal Records Council and National Archives and Records Administration 2014 Federal RIM Program Maturity Model User's Guide. Available at https://www.archives.gov/records-mgmt/prmd.html [Last accessed 29 April 2016]

Office of Management and Budget 2013 Open Data Policy – Managing Information as an Asset. Washington, DC. Available at https://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf [Last accessed 29 April 2016].

Office of Science and Technology Policy 2013 Increasing Access to the Results of Federally Funded Scientific Research. Washington, DC. Available at https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf [Last accessed 29 April 2016].

Peng G, Privette J, Kearns E, Ritchey N, and Ansari S 2015 A UNIFIED FRAMEWORK FOR MEASURING STEWARDSHIP PRACTICES APPLIED TO DIGITAL ENVIRONMENTAL DATASETS. In Data Science Journal, Volume 13, 2 February 2015. Available at http://datascience.codata.org/articles/abstract/10.2481/dsj.14-049/ [Last accessed 29 April 2016].

Phillips M, Bailey J, Goethals A, and Owens T 2013 The NDSA Levels of Digital Preservation: An Explanation and Uses. Available at http://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf [Last accessed 2 May 2016].

Whyte, A 2015 Where to keep research data: DCC Checklist for Evaluating Data Repositories. v.1 Edinburgh: Digital Curation Centre. Available at http://www.dcc.ac.uk/resources/how-guides [Last accessed 29 April 2016].